
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA
Prompting in Vision
Towards conversational visual intelligence
19 June 2023, 9AM-12PM
https://prompting-in-vision.github.io/
Image: unsplash.com/@jean_vella

Prompting in Vision

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

https://prompting-in-vision.github.io/

Kaiyang Zhou
NTU

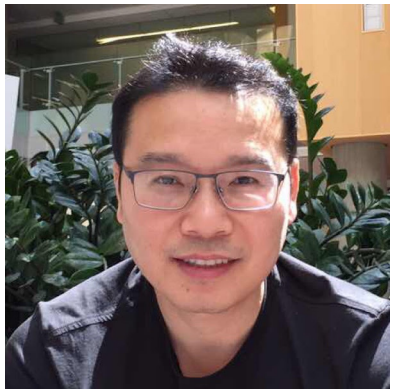
Ziwei Liu
NTU

Phillip Isola
MIT

Hyojin Bahng
MIT

Ludwig Schmidt
U. of Washington

Sarah Pratt
U. of Washington



Denny Zhou
Google Brain

Schedule

09:10 am - 09:50 am Prompting in visual intelligence and generation
09:50 am - 10:30 am Teaching language models to reason
10:30 am - 10:40 am Coffee break
10:40 am - 11:20 am Visual prompting
11:20 am - 12:00 pm Improved model adaptation via weight interpolation and prompt generation

unsplash.com/@chris23

Tutorial 1: Prompting in visual intelligence and generation

09:10 am – 09:50 am

Kaiyang Zhou

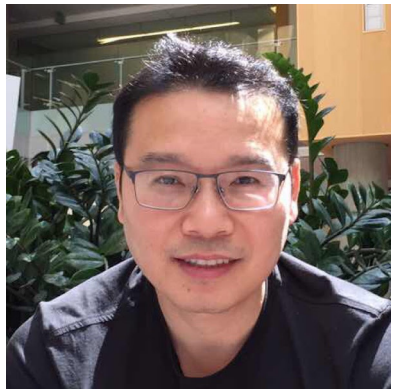
- Postdoc at Nanyang Technological University
- Incoming Assistant Professor at Hong Kong Baptist University
- PhD (2020) from the University of Surrey
- Open-world visual learning (generalization, foundation models)
- https://kaiyangzhou.github.io/

Ziwei Liu

- Assistant Professor at Nanyang Technological University
- PhD (2017) from the Chinese University of Hong Kong
- Vision, learning, and graphics
- Burst Denoising, CelebA, DeepFashion, Fashion Landmarks, DeepMRF, Voxel Flow, Long Tail, Compound Domain, and Wildlife Conservation
- https://liuziwei7.github.io/

Tutorial 2: Teaching language models to reason

09:50 am - 10:30 am



Denny Zhou

- Founder and lead of the Reasoning Team in Google Brain & DeepMind
- Build and teach LLMs to achieve human-level reasoning
- Instruction tuning (FLAN2), chain-of-thought prompting, self-consistency decoding, least-to-most prompting, and emergent properties of LLMs
- https://dennyzhou.github.io/

Tutorial 3: Visual prompting

10:40 am - 11:20 am

Phillip Isola

- Associate Professor in EECS at MIT
- PhD in Brain & Cognitive Sciences at MIT
- Science of intelligence: science of deep learning, emergent intelligence, embodied intelligence, role of data and environments, and controllable AI
- pix2pix, CycleGAN
- http://web.mit.edu/phillipi/

Hyojin Bahng

- PhD student at MIT CSAIL
- Master from Korea University
- Computer vision and machine learning
- https://hjbahng.github.io/

Tutorial 4: Improved model adaptation via weight interpolation and prompt generation

11:20 am - 12:00 pm

Ludwig Schmidt

- Assistant Professor at the University of Washington
- Postdoc at UC Berkeley and PhD from MIT
- Foundations of machine learning (datasets, evaluation, reliable generalization, and large models)
- OpenCLIP, OpenFlamingo, and LAION-5B
- https://people.csail.mit.edu/ludwigs/

Sarah Pratt

- PhD student at the University of Washington
- Previous member of the PRIOR team at The Allen Institute for AI
- Machine learning and computer vision
- https://sarahpratt.github.io/

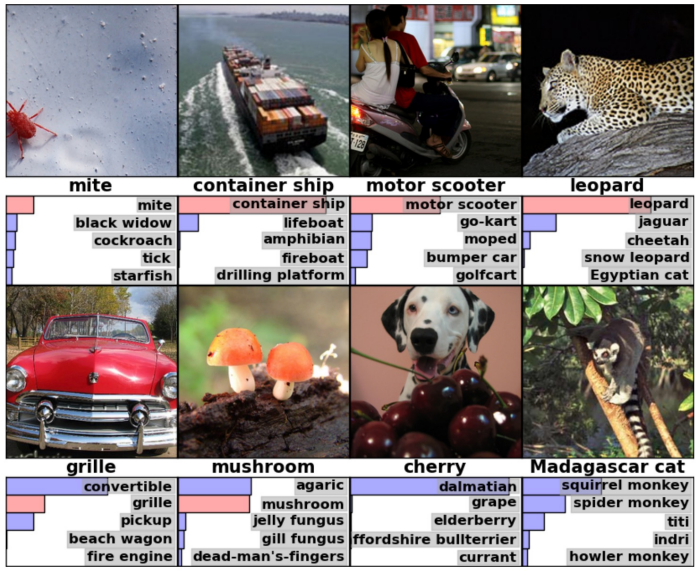
Prompting in Visual Intelligence

Kaiyang Zhou
Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

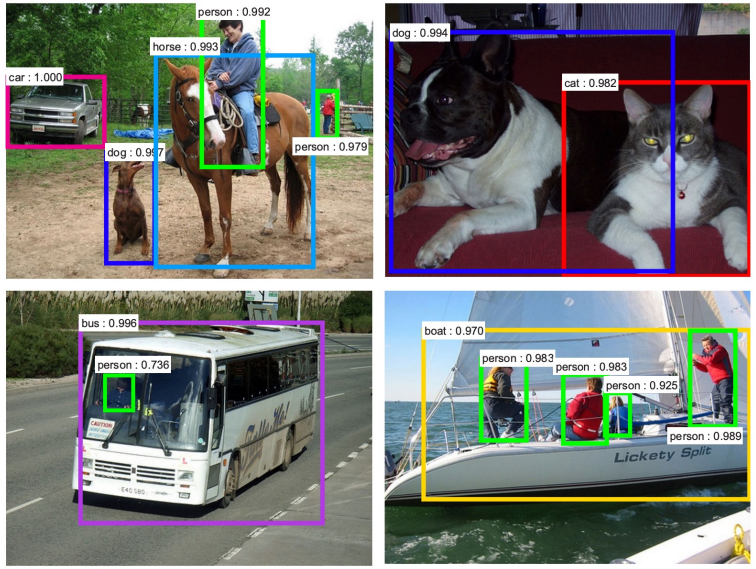
10 years ago

Classification



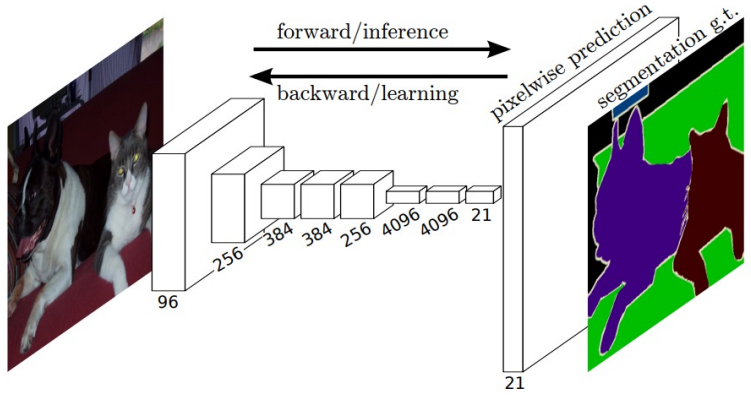
Krizhevsky et al., 2012

Detection



Ren et al., 2015

Segmentation



Long et al., 2015

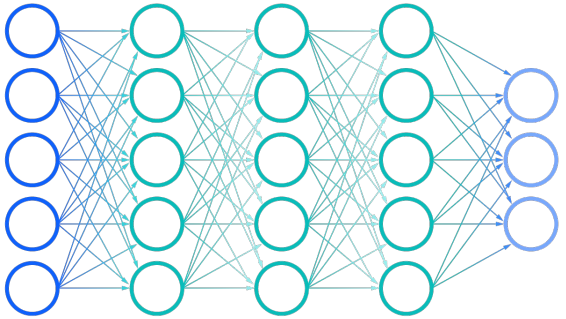
...

10 years ago

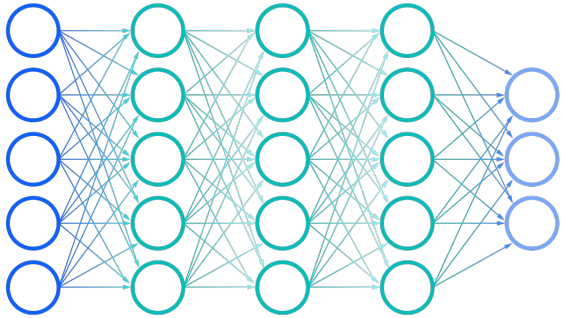
1. Specify targets
2. Collect a dataset
3. Train a neural net

Development stage

Dog
Cat
Bird



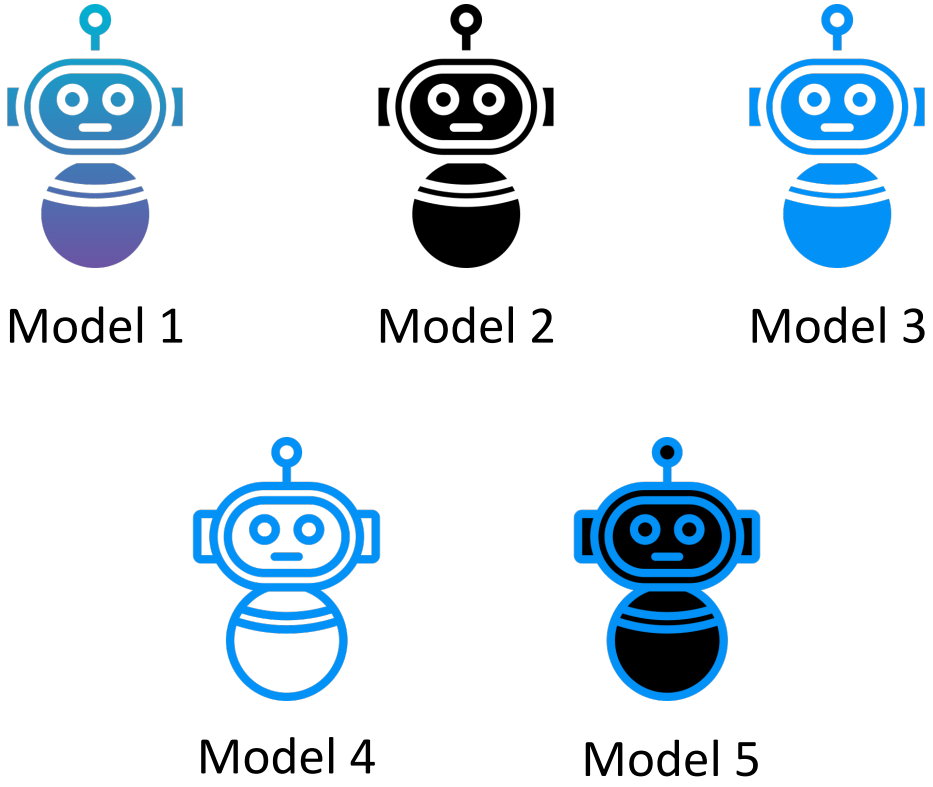
Deployment stage



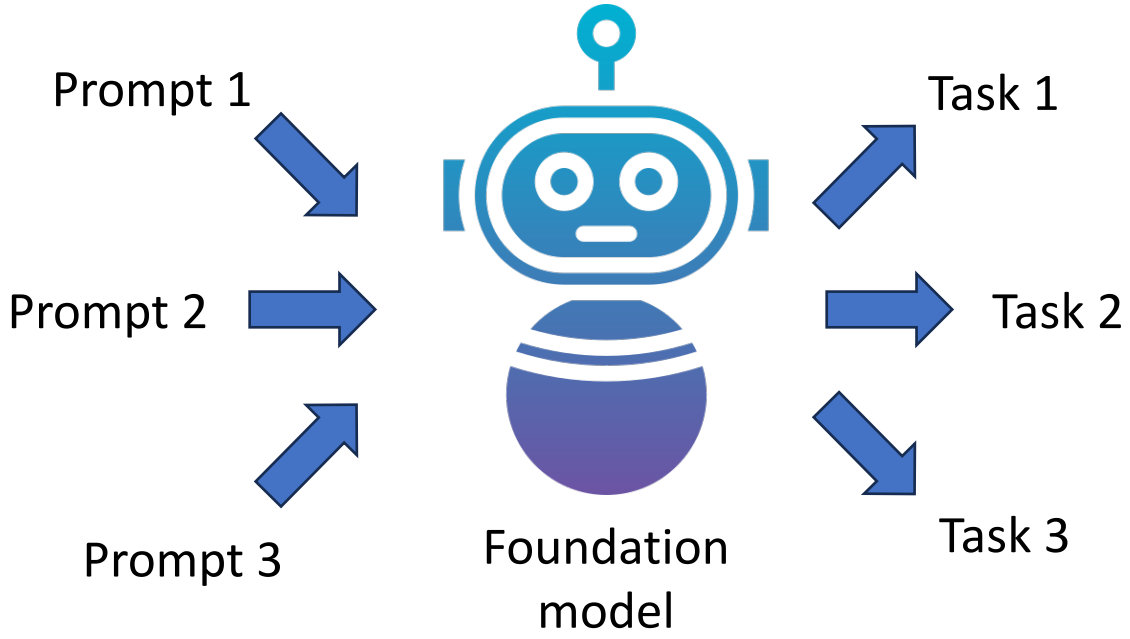
Dog
Cat
Bird

2013 vs. 2023

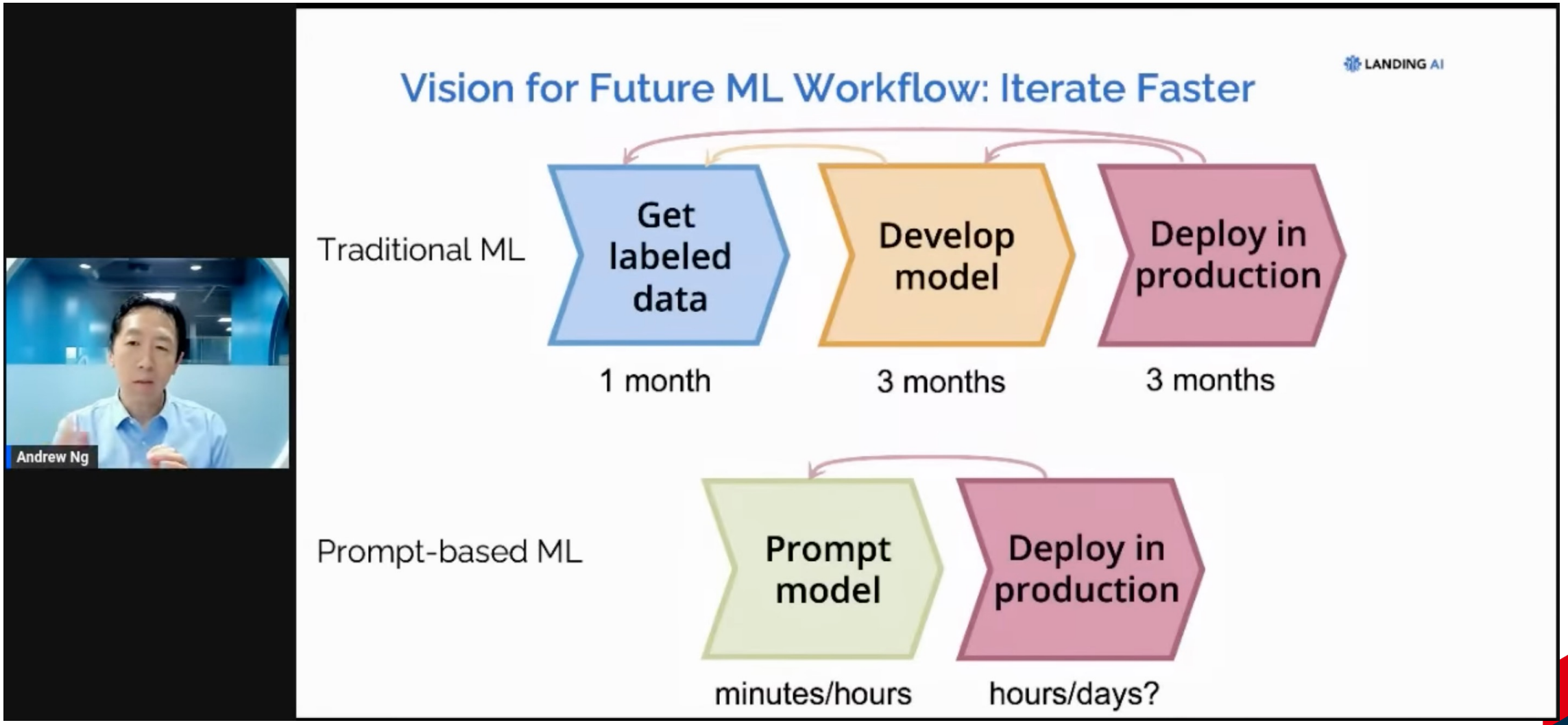
Old days: one model for one purpose



Now: one model for multiple purposes



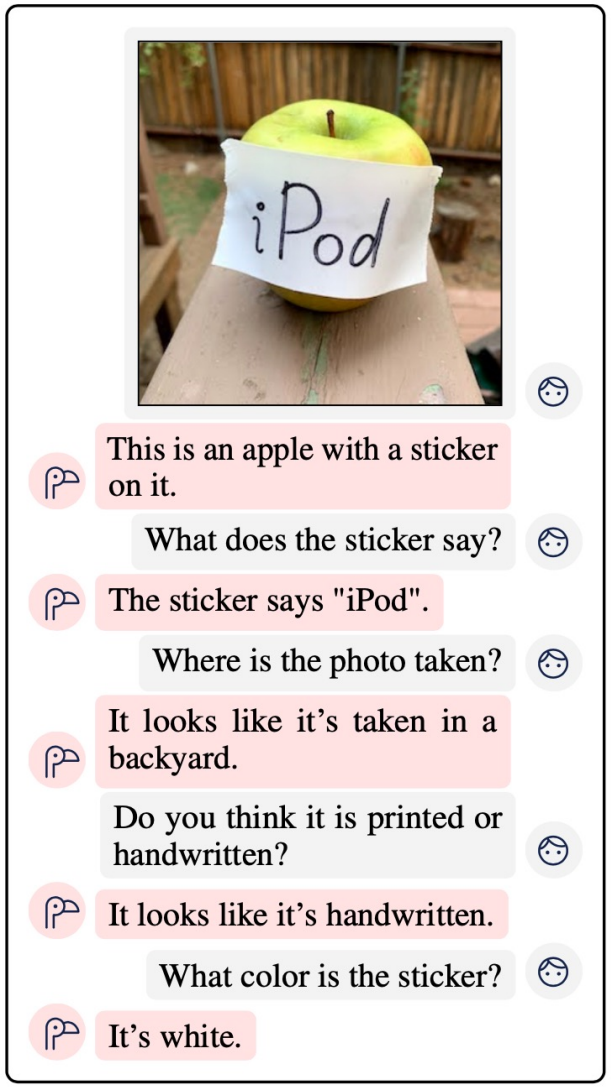
Andrew Ng: Visual prompting is an emerging path …



Credit: Visual Prompting Livestream With Andrew Ng, Landing AI

2023: Prompting is everywhere

DeepMind's Flamingo



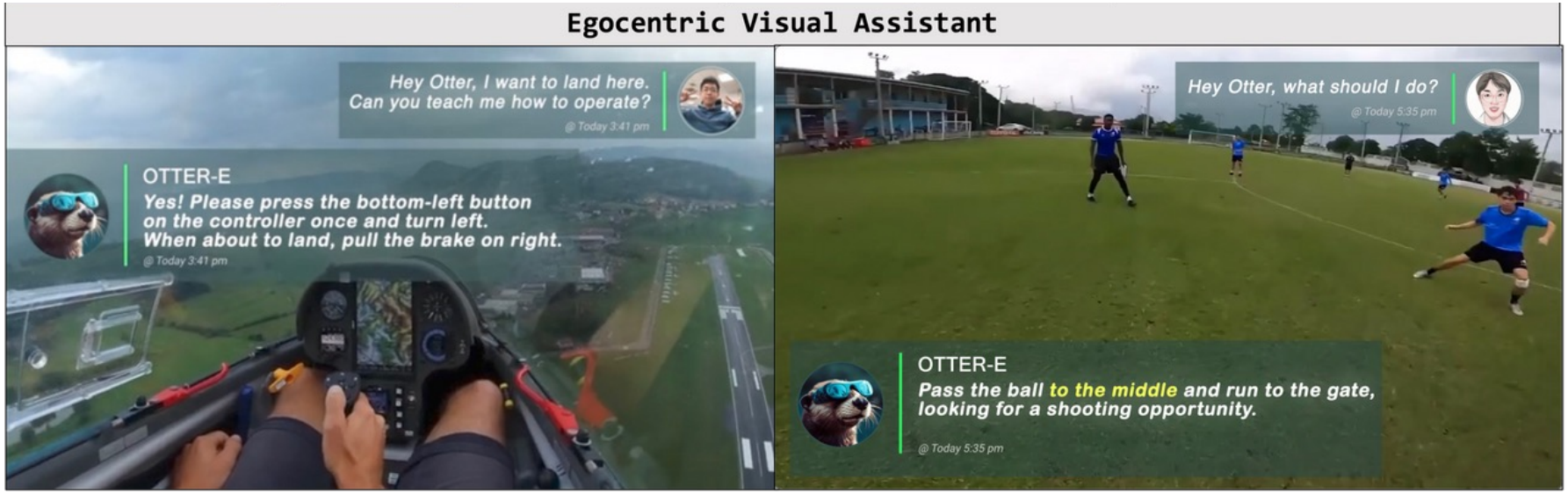
Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS'22.

2023: Prompting is everywhere

OTTER

A Multi-Modal Model with In-Context Instruction Tuning

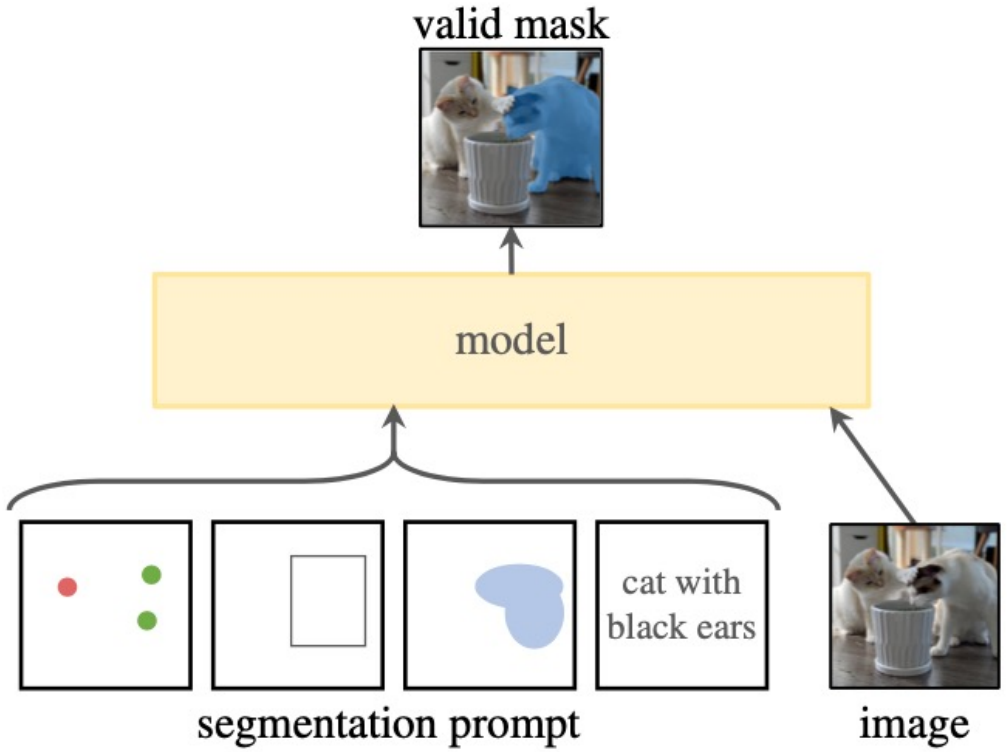
Bo Li*[1] Yuanhan Zhang*,[1] Liangyu Chen*,[1] Jinghao Wang*,[1] Fanyi Pu*,[1]
Jingkang Yang[1] Chunyuan Li[2] Ziwei Liu[1]
[1]S-Lab, Nanyang Technological University [2]Microsoft Research, Redmond



Li et al. Otter: A Multi-Modal Model with In-Context Instruction Tuning. arXiv:2305.03726.

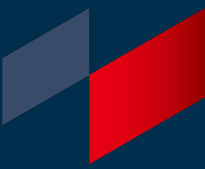
2023: Prompting is everywhere

Meta AI' SAM model

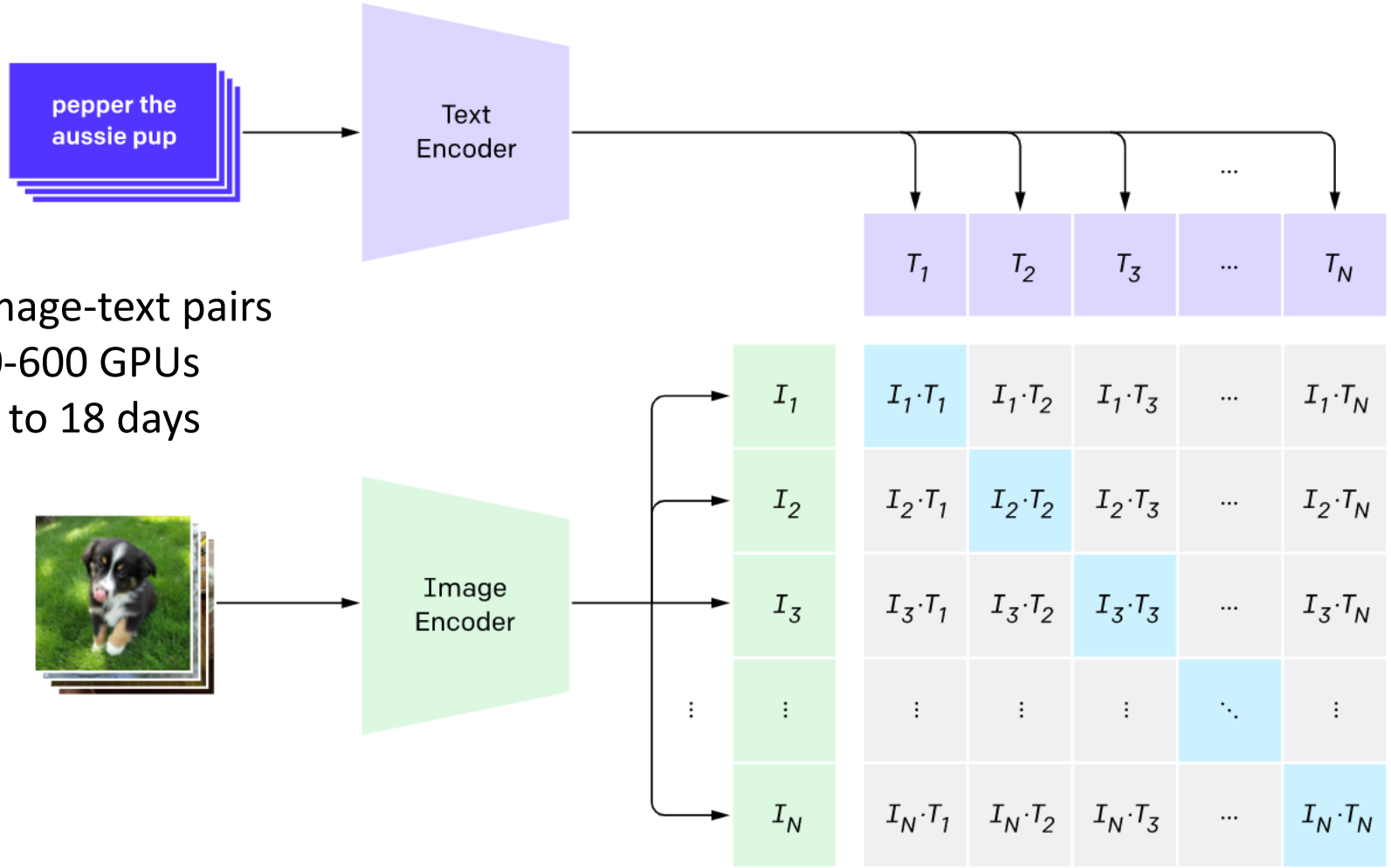


Kirillov et al. Segment Anything. arXiv:2304.02643.

Prompt learning for visual language models



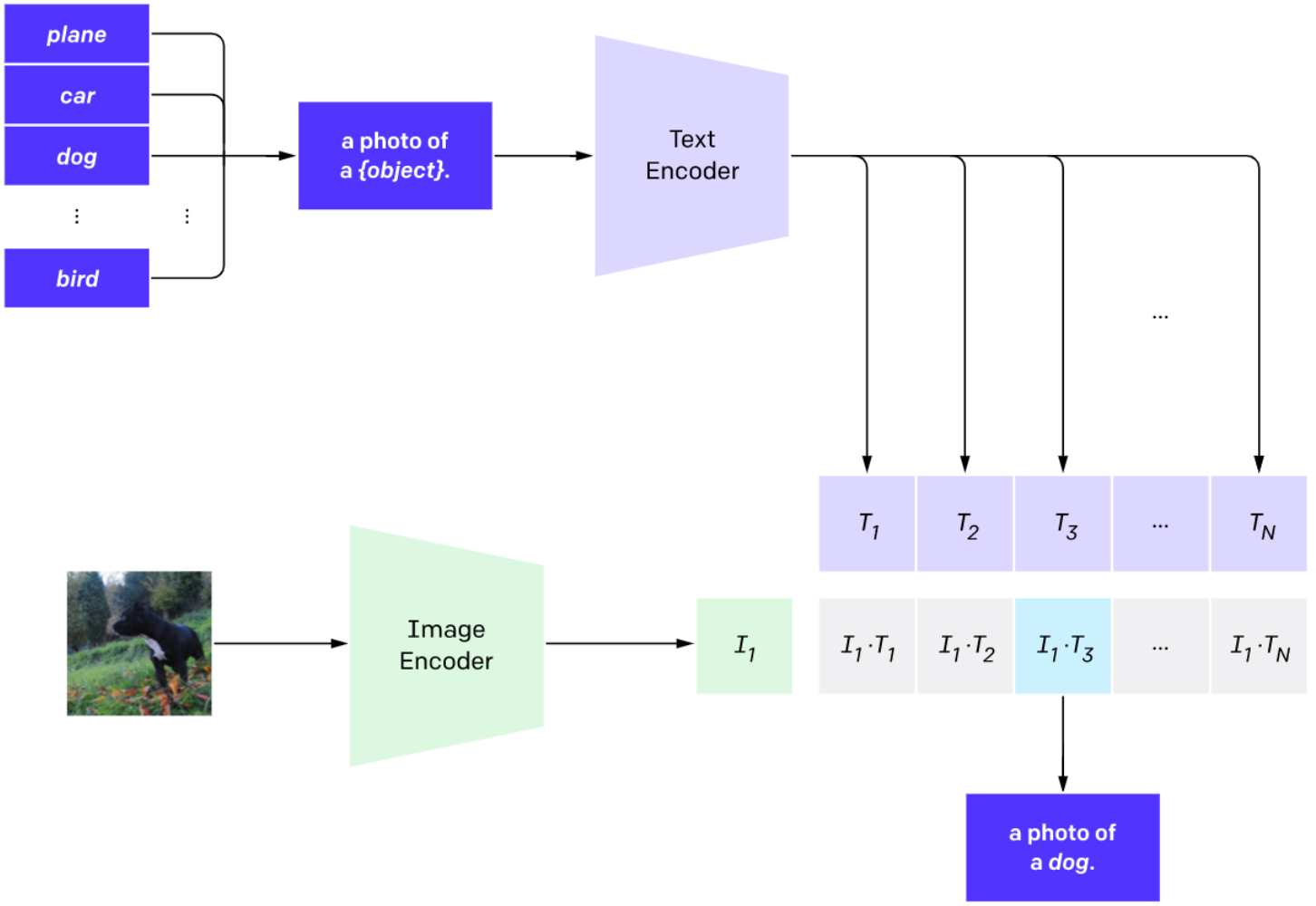
Contrastive Language-Image Pre-training (CLIP)



- Data: 400M image-text pairs
- Compute: 250-600 GPUs
- Train time: up to 18 days

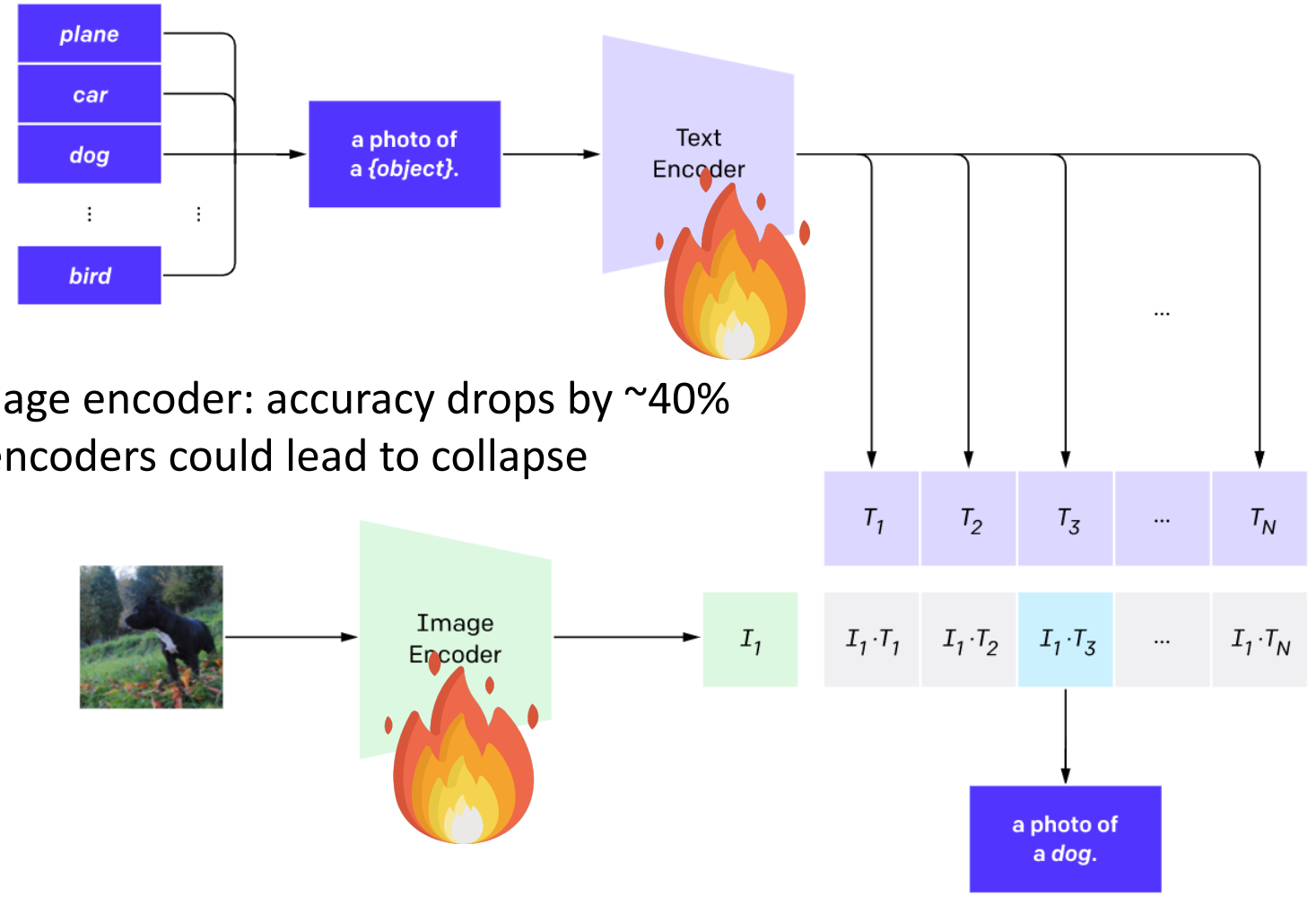
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

Zero-shot image recognition via prompting



Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

Fine-tuning might not be a good idea



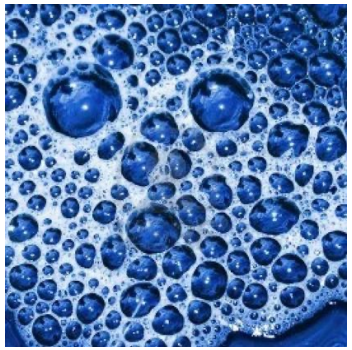
- Fine-tuning the image encoder: accuracy drops by ~40%
- Fine-tuning both encoders could lead to collapse

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

Prompt engineering is too time-consuming

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**91.83**

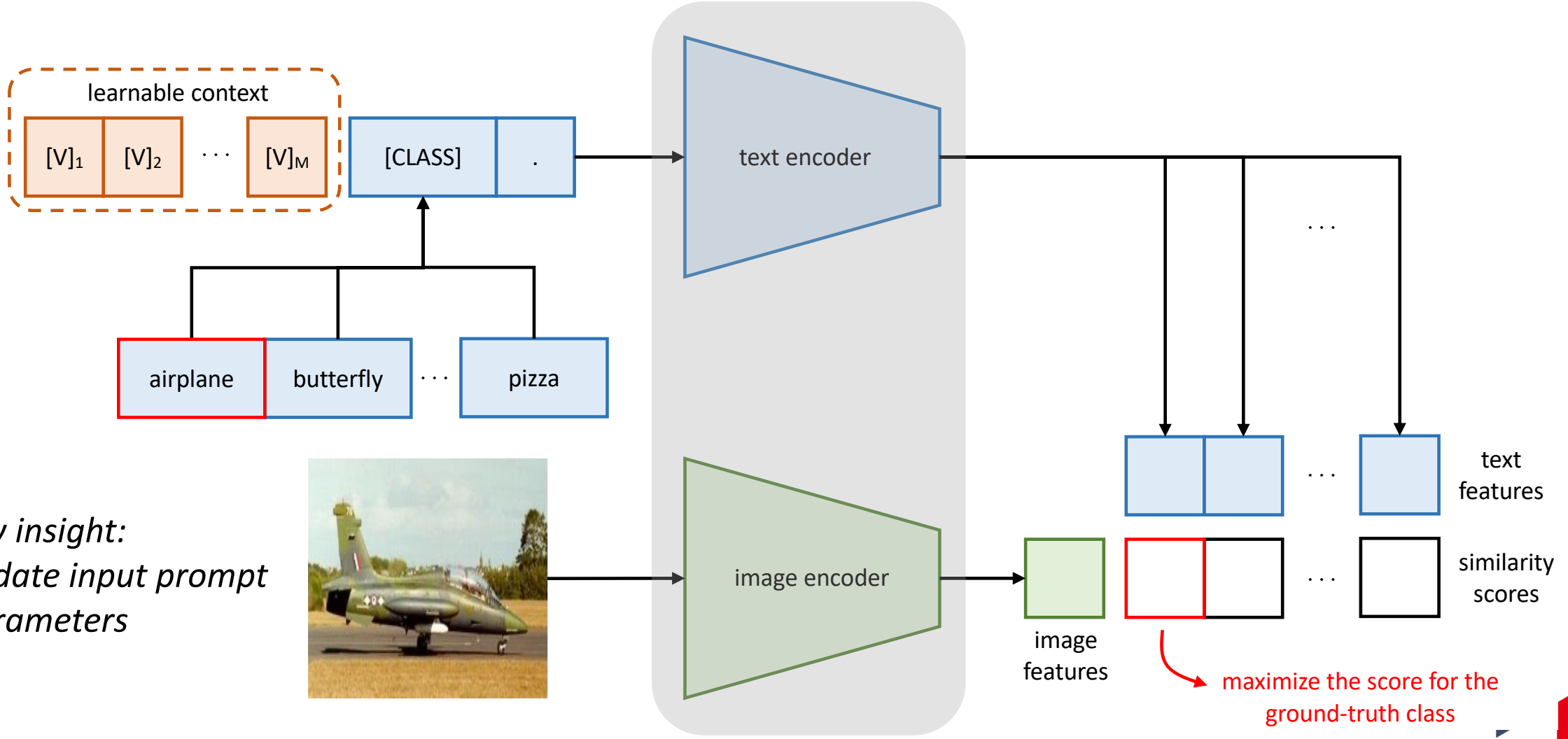
Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**94.51**

Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**63.58**

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**83.53**

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

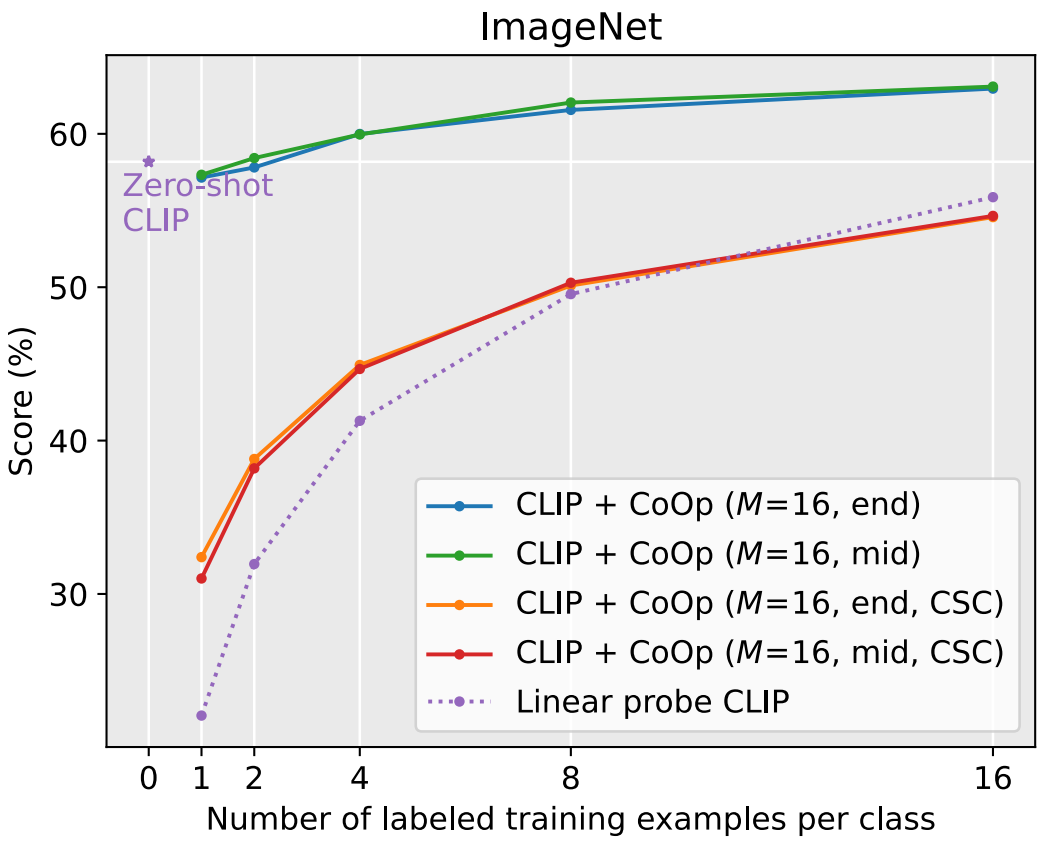
Context Optimization (CoOp)

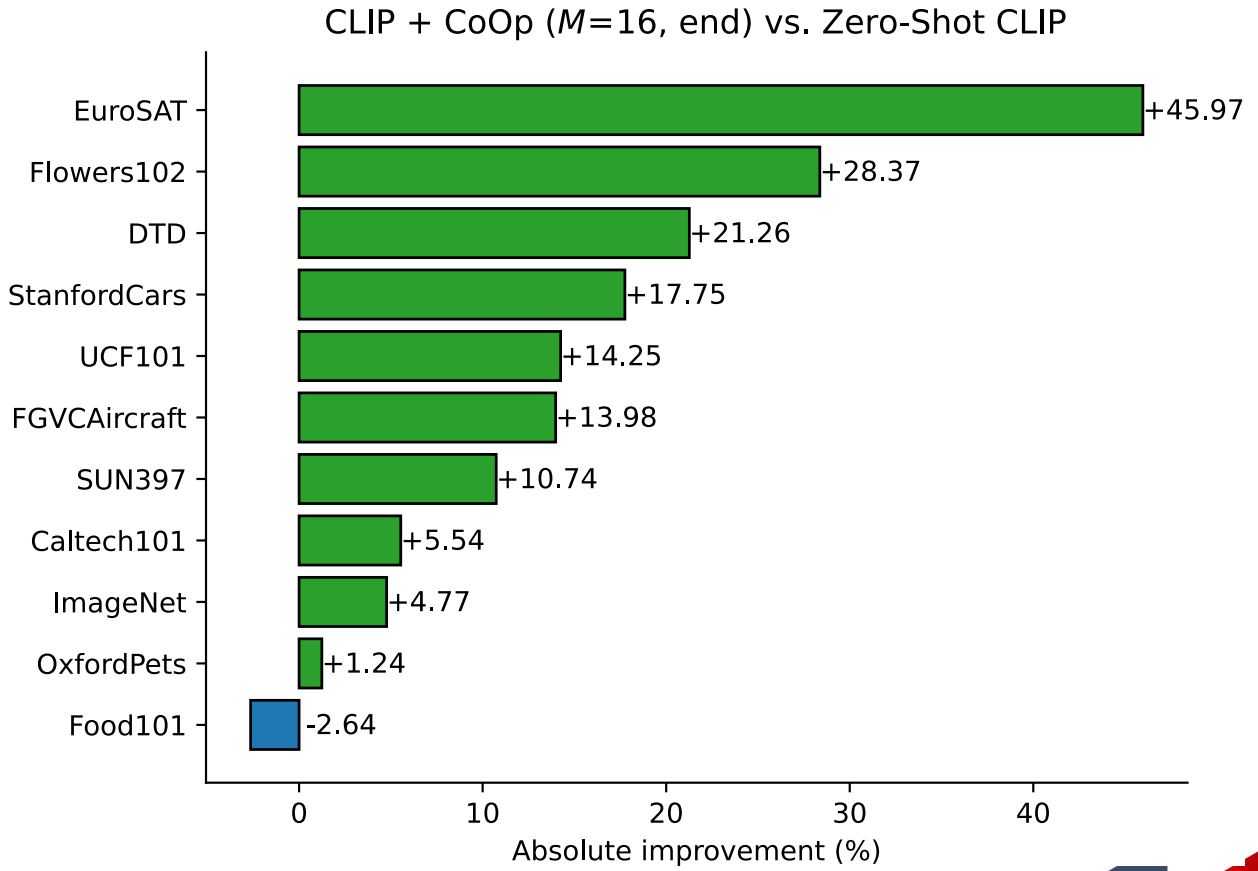


Key insight: update input prompt parameters

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

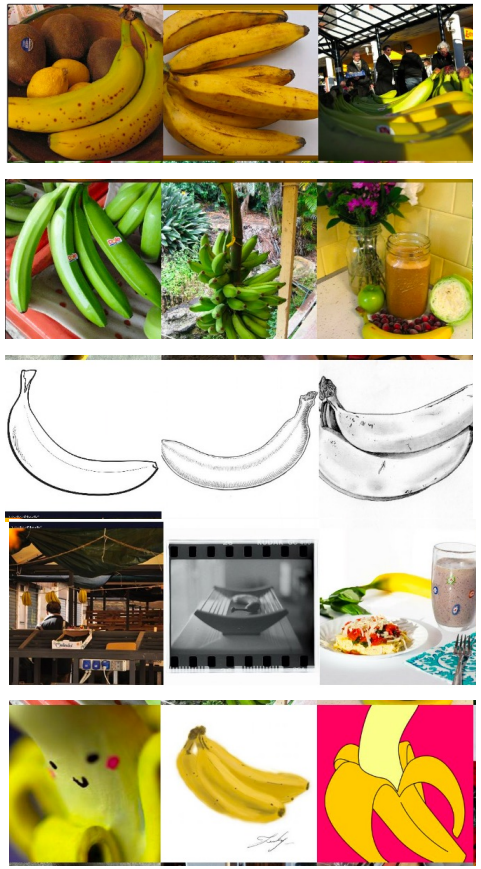
CoOp is a few-shot learner





Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

CoOp is domain-generalizable



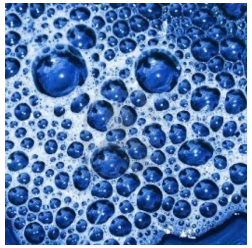
	CLIP	Ours
ImageNet (source)	58.18	**63.33**
V2 (target)	51.34	**55.40**
Sketch (target)	33.32	**34.67**
Adversarial (target)	21.65	**23.06**
Rendition (target)	56.00	**56.60**

Method	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M=16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M=4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M=16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M=4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M=16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M=4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M=16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M=4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

Visualization




A few words are somewhat relevant

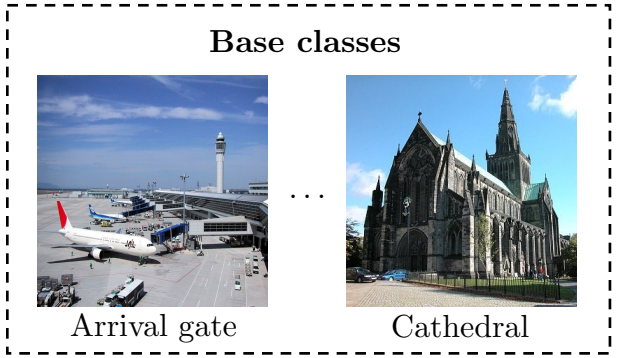
But the whole prompt doesn't make much sense

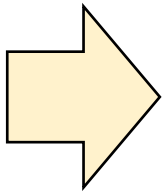
#	ImageNet	Food101	OxfordPets	DTD	UCF101
1	Potd (1.7136)	Lc (0.6752)	Tosc (2.5952)	Boxed (0.9433)	Meteorologist (1.5377)
2	That (1.4015)	Enjoyed (0.5305)	Judge (1.2635)	Seed (1.0498)	Exe (0.9807)
3	Filmed (1.2275)	Beh (0.5390)	Fluffy (1.6099)	Anna (0.8127)	Parents (1.0654)
4	Fruit (1.4864)	Matches (0.5646)	Cart (1.3958)	Mountain (0.9509)	Masterful (0.9528)
5	,... (1.5863)	Nytimes (0.6993)	Harlan (2.2948)	Eldest (0.7111)	Fe (1.3574)
6	°(1.7502)	Prou (0.5905)	Paw (1.3055)	Pretty (0.8762)	Thof (1.2841)
7	Excluded (1.2355)	Lower (0.5390)	Incase (1.2215)	Faces (0.7872)	Where (0.9705)
8	Cold (1.4654)	N/A	Bie (1.5454)	Honey (1.8414)	Kristen (1.1921)
9	Stery (1.6085)	Minute (0.5672)	Snuggle (1.1578)	Series (1.6680)	Imam (1.1297)
10	Warri (1.3055)	~ (0.5529)	Along (1.8298)	Coca (1.5571)	Near (0.8942)
11	Marvelcomics (1.5638)	Well (0.5659)	Enjoyment (2.3495)	Moon (1.2775)	Tummy (1.4303)
12	.: (1.7387)	Ends (0.6113)	Jt (1.3726)	Ih (1.0382)	Hel (0.7644)
13	N/A	Mis (0.5826)	Improving (1.3198)	Won (0.9314)	Boop (1.0491)
14	Lation (1.5015)	Somethin (0.6041)	Srsly (1.6759)	Replied (1.1429)	N/A
15	Muh (1.4985)	Seminar (0.5274)	Asteroid (1.3395)	Sent (1.3173)	Facial (1.4452)
16	.# (1.9340)	N/A	N/A	Piedmont (1.5198)	During (1.1755)

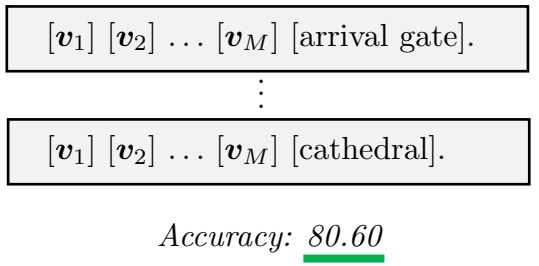
Takeaway:
- Difficult to interpret continuous vectors using the nearest words approach
- The concepts learned in those vectors probably go beyond the vocabulary
- Need to trade-off between performance and interpretability

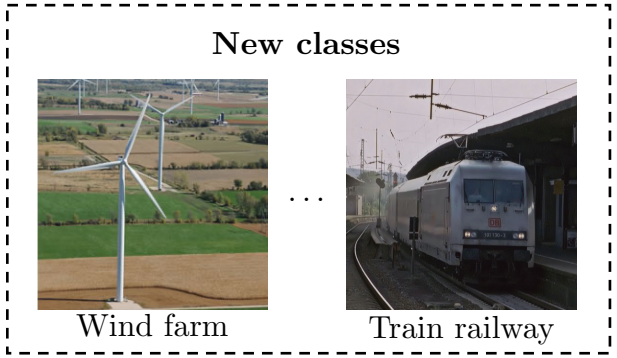
Zhou, Yang, Loy, and Liu. Learning to Prompt for Vision-Language Models. IJCV'22.

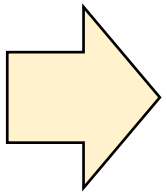
Can CoOp generalize to broader concepts within the same dataset?

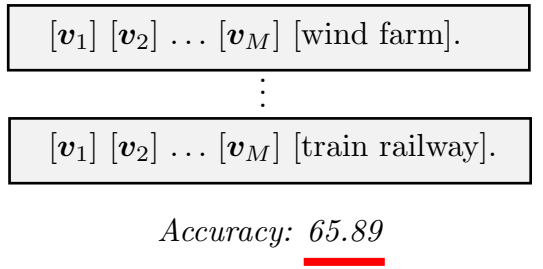












Problem: The prompt only works for a subset of classes (i.e., overfitting)

Zhou, Yang, Loy, and Liu. Conditional Prompt Learning for Vision-Language Models. CVPR'22.

More failure cases of CoOp on unseen classes

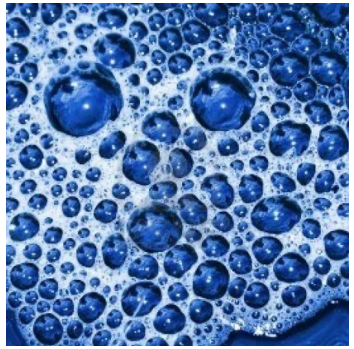
ImageNet
↓8.86%

Caltech101
↓8.19%

Flowers102
↓37.93%

StanfordCars
↓17.72%

FGVCAircraft
↓18.14%



DTD
↓38.26%

EuroSAT
↓37.45%

UCF101
↓28.64%

Zhou, Yang, Loy, and Liu. Conditional Prompt Learning for Vision-Language Models. CVPR'22.

What is a good prompt?

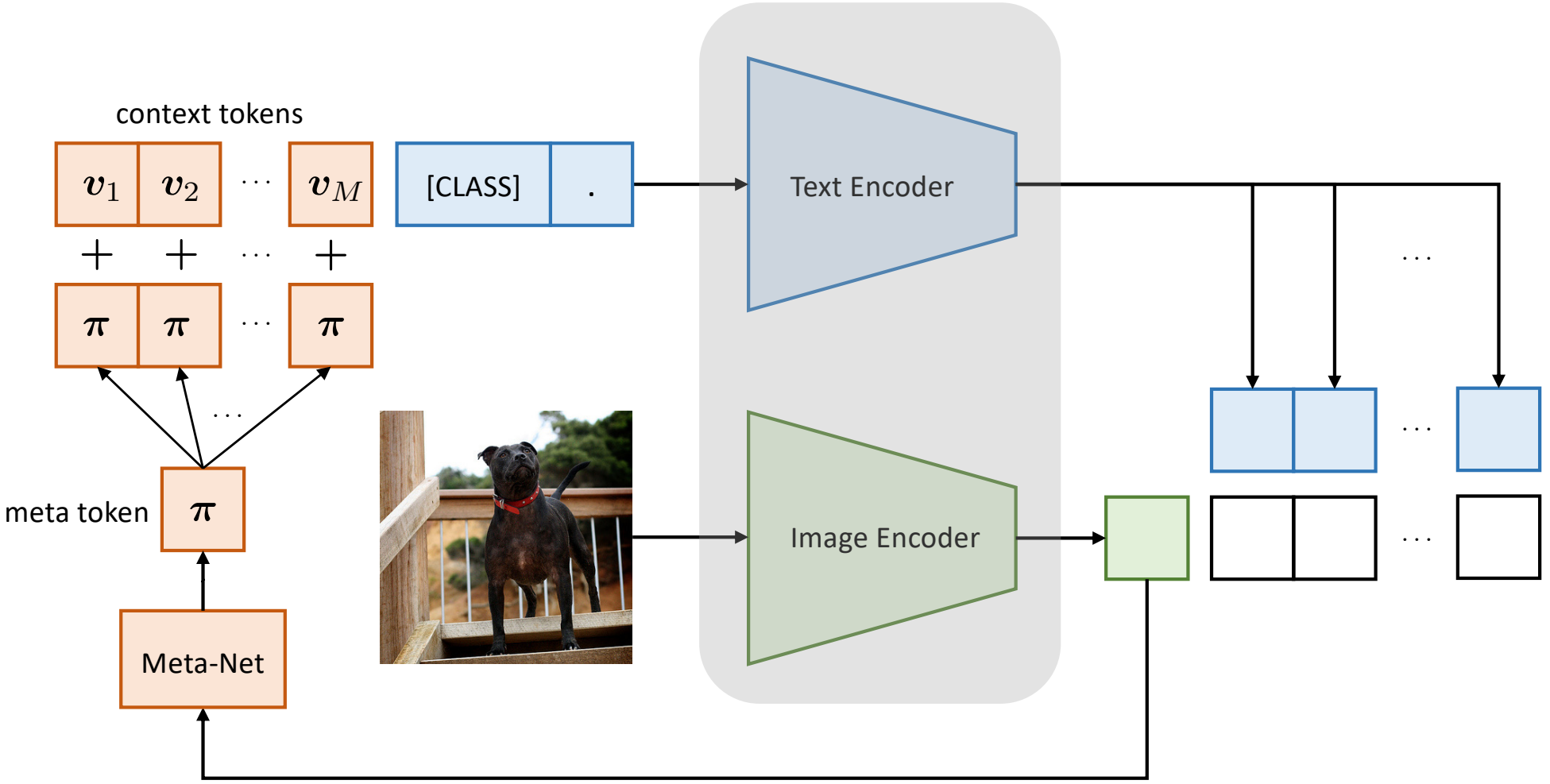
A good prompt should characterize each instance with some specific context words

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

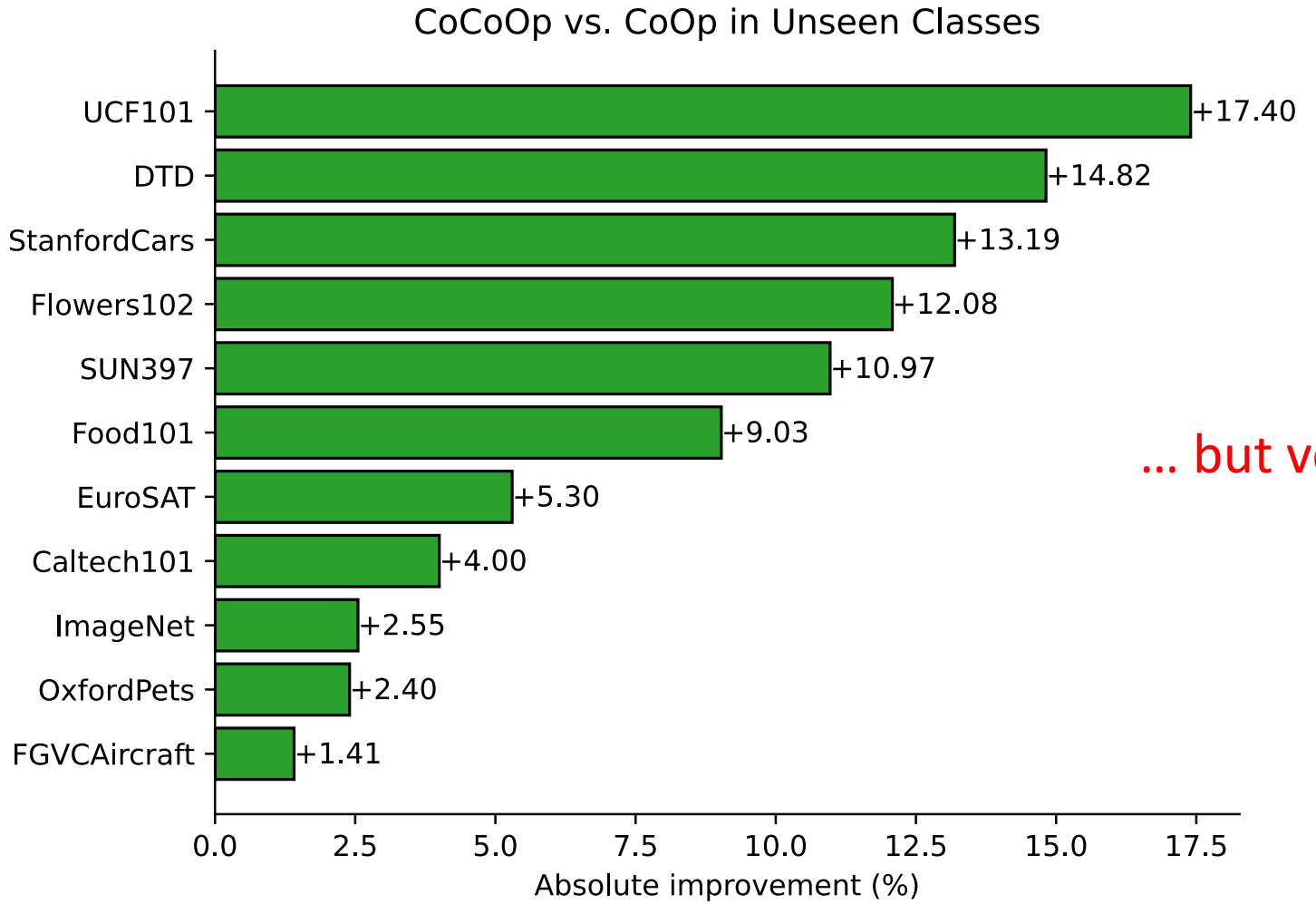
Zhou, Yang, Loy, and Liu. Conditional Prompt Learning for Vision-Language Models. CVPR'22.

Conditional Context Optimization (CoCoOp)



Zhou, Yang, Loy, and Liu. Conditional Prompt Learning for Vision-Language Models. CVPR'22.

Conditional prompt learning is more generalizable



… but very slow to train

Zhou, Yang, Loy, and Liu. Conditional Prompt Learning for Vision-Language Models. CVPR'22.

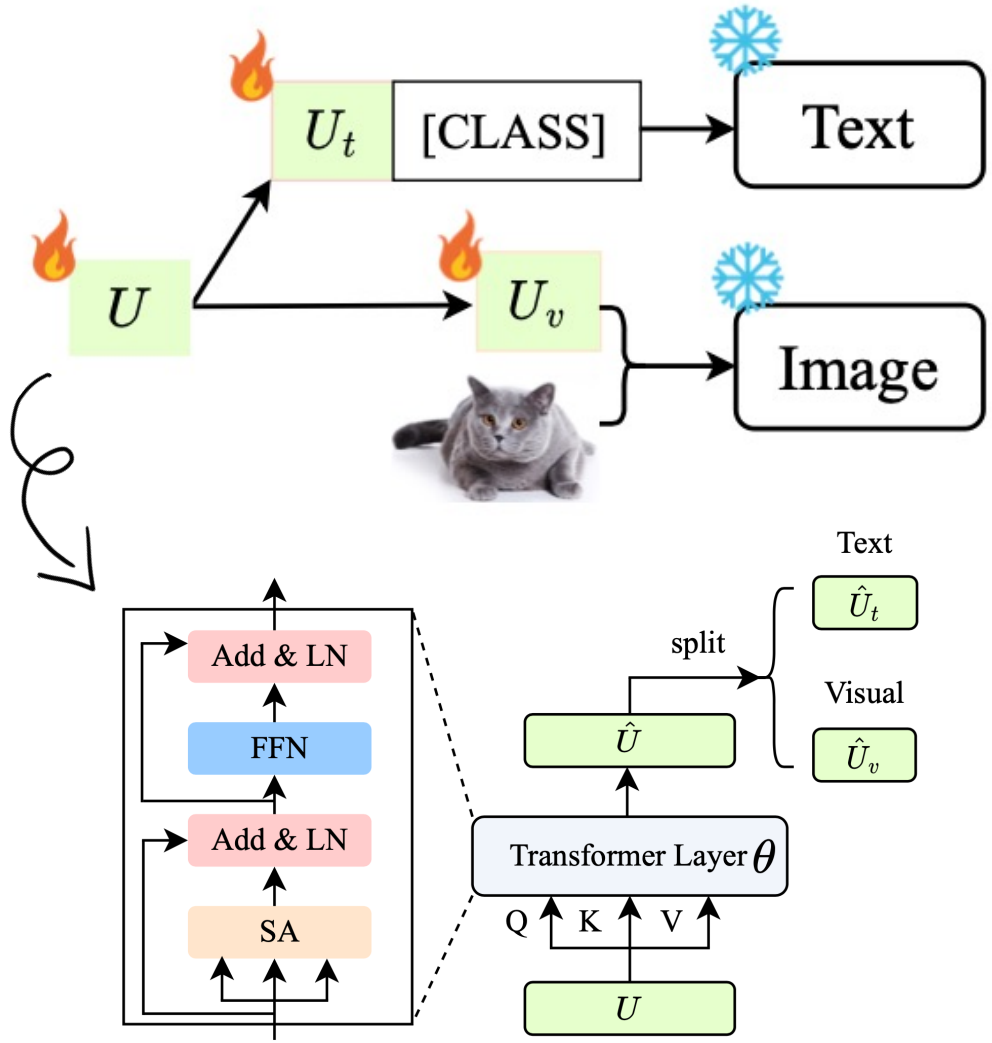
Multimodal prompt learning is more efficient while performs similarly well

$$\boldsymbol{w}^T\boldsymbol{x}$$

Text features

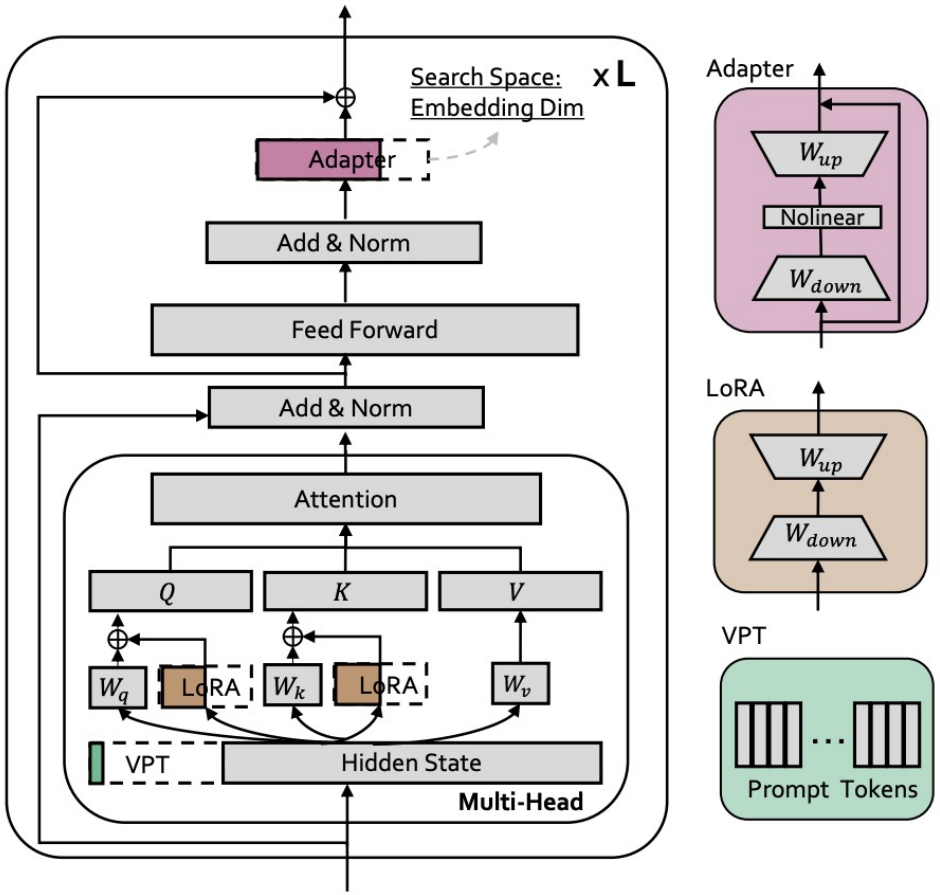
Image features

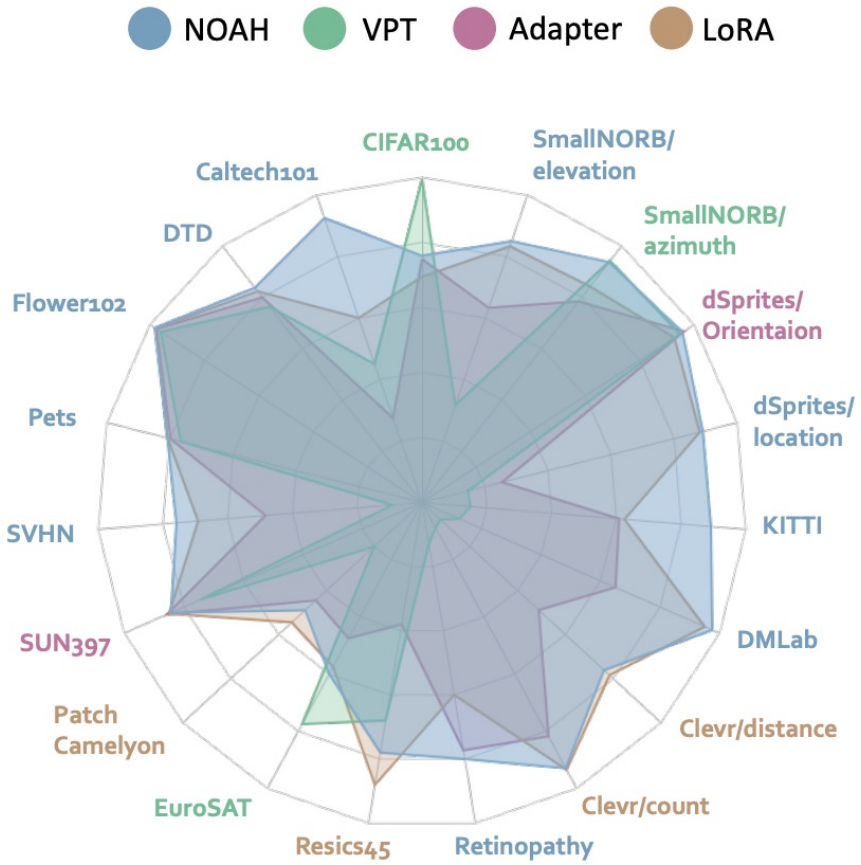
Tune the learnable parameters in both input spaces



Zang, Li, Zhou, Huang, and Loy. Unified Vision and Language Prompt Learning. arXiv:2210.07225.

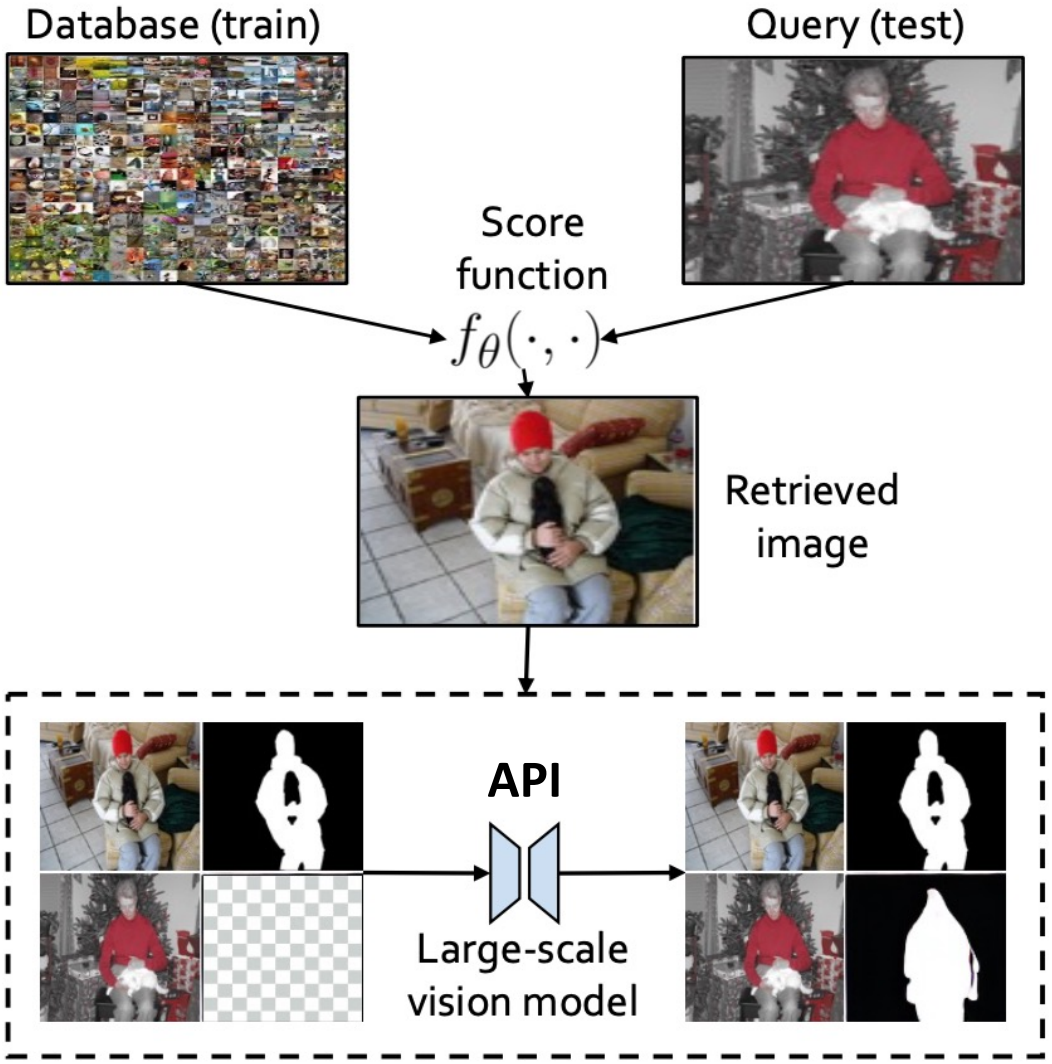
Do neural prompt search if you have more compute resources





Zhang, Zhou, and Liu. Neural Prompt Search. arXiv:2206.04673.

Black-box prompt learning: prompt retrieval



$$x^* = \arg\max_{x_n \in \mathcal{D}} f_\theta(x_n, x_q)$$

Use the API's output as supervision signal

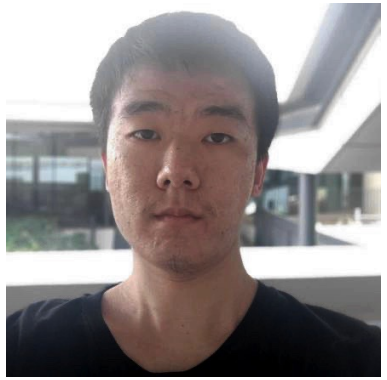
Zhang, Zhou, and Liu. What Makes Good Examples for Visual In-Context Learning? arXiv:2301.13670.

Take-home messages

- Prompt learning is data-efficient (allows few-shot learning)
- Conditional prompt learning is more generalizable (but takes much more compute)
- Multimodal prompt learning offers better trade-offs between performance and training time
- Have more compute resources? Do neural prompt search for the entire model
- Only have access to APIs? Use prompt retrieval

Acknowledgement


Jingkang Yang


Yuhang Zang


Yuanhan Zhang


Ziwei Liu


Chen Change Loy

References

- Learning to Prompt for Vision-Language Models. IJCV'22.
- Conditional Prompt Learning for Vision-Language Models. CVPR'22.
- Unified Vision and Language Prompt Learning. arXiv:2210.07225.
- Neural Prompt Search. arXiv:2206.04673.
- What Makes Good Examples for Visual In-Context Learning? arXiv:2301.13670.

Code: https://github.com/KaiyangZhou

Paper pdfs: https://kaiyangzhou.github.io/pub.html